Chapter 5

Planning and Implementation

By this time you will have done an assessment of what you need to preserve and your available resources, developed goals for your digital preservation program, written a digital preservation policy, and engaged your stakeholders to get the program off the ground. Now you need to decide how to implement a digital preservation program that meets your institution's needs but is also within your organization's abilities to develop and sustain. A common reference for planning a digital preservation program is the digital curation life cycle model (figure 5.1).

The Digital Curation Centre (DCC) Curation Lifecyle Model introduces a visualization of the concept that digital preservation is an iterative process that builds, changes, and improves through each cycle of maintenance.[1] The core of the model, which is a series of concentric circles, is what the preservation program is trying to protect and provide access to, the information packages of digital content and metadata. Moving outward from the center, the model introduces the staged manner in which practitioners should approach building a preservation program. Before anything else and throughout the entire life of the preservation program, preservation planning is paramount. The next level out is community watch and participation, which is an integral piece of preservation planning. Moving further away from the center, the model provides the first level of granularity, where curation and preservation become distinct parts of the preservation program. Finally, curation and preservation are broken down into actionable and sequential stages of a preservation program. The parts of the model that are not in the circle are actions that occur only after a triggering event, such as a change to collection development policy that would require you to determine if some of the materials maintained by your digital preservation program may no longer be within the scope of your collection. Another triggering event may be that you have a large collection of photographs in a format that is no longer accessible. This would require you to migrate the data into a new format so that you can provide uninterrupted access of the material to your users.

When using the life cycle model to plan your digital preservation program, there are two common pathways to follow. The first pathway is to decide that you would prefer to subscribe to an out-of-the-box digital preservation vendor that provides a system that covers all aspects of the preservation life cycle from "create or receive" to "transform." There are several of these vendors on the market, with about an even split of proprietary systems and hosted solutions that package together open-source tools and preservation storage systems. The other pathway to designing your digital preservation program is to develop your own system, using a series of open-source and commercial tools. Which pathway you choose will depend on the resources you have available in your organization and how far along you are in your preservation program.

The benefits of the vendor solutions are that you will only need to learn how to use one software system that integrates all aspects of the preservation life cycle and that the system is maintained and updated by an external party. If your personnel resources and expertise are limited, this is a great solution to implementing an efficient and sustainable program. These solutions can be expensive, and, depending on the size of your organization, duplicate systems may already be available and implemented by your information technology department (particularly storage systems). Like other specialized systems, proprietary vendor solutions are a small market, and it could be difficult to replace one if something goes wrong or to exit from one if you are unsatisfied with its services.[2]

The second pathway toward building a digital preservation program allows you complete flexibility and the ability to build your system one piece at a time as your resources for and knowledge about digital preservation increases. Due to the foundational values of digital preservation being sustainability and collaboration, most of the necessary tools for building and maintaining a digital preservation program are open-source and are maintained by a dedicated community

of practitioners. It is entirely possible to build a digital preservation program using only free, open-source tools, but that requires that you have someone on staff who has the time to devote to the program and who has a high level of technology competency.

The Preserving Digital Objects with Restricted Resources (Digital POWRR) project created a tool grid in 2013 that compared digital preservation tools using categories drawn from the OAIS Reference Model. This grid has not been updated since it was created, but it provides a snapshot of the most commonly used digital preservation tools and their suitability for different aspects of the digital preservation life cycle. The most up-to-date listing of tools is the Community Owned Digital Preservation Tool Registry (COPTR), which has recently added a subsite of Community Owned Workflows (COW).[3] These resources are essential when initially planning your digital preservation program because they can save you time. The comparison between different tools has already been done, and all you have to decide is which tools work best for your particular situation.

Before implementing your digital preservation program, either a full end-to-end system or a patchwork of tools and services, decide on who will have the authority to access and, when necessary, modify your digital content at each stage of the preservation life cycle. Establishing a transparent, trustworthy, and secure digital preservation program requires a clear set of authority controls—who has permission to read, write, modify, and delete digital content in your preservation system. In the beginning stages of the preservation life cycle, very few people should have access to the content to prevent accidental or malicious alterations. As the materials move through the preservation workflow, there will be two sets of permissions needed—one set for users accessing the fully arranged and described content, and one set for those managing the digital preservation master files, or archival information packages.[4] The permissions for researchers will depend upon the collection and your institution's rules. There are specialty repository systems, such as Mukurtu, where access restrictions can be set to conform to cultural practice.[5] The permissions for the preservation masters should be limited to the specific person or people in your information technology division and the archivists responsible for maintaining your digital preservation program. Also, in the case of the preservation masters, you should have clear guidelines that state *when* it is permissible to access these masters and for what purposes. The key to a trusted digital preservation program is transparency and consistency. Whatever pathway you take, document all of your workflows and procedures, consistently perform procedures as documented, and record any changes you make to your workflows, including why the changes were made, when, and by whom.

The digital preservation program that I manage is based upon the second pathway, a series of workflows using open-source and commercial tools and systems. The overall flow of the program is stabilize; appraise, arrange, and describe; ingest; access. This is a little different from the digital curation life cycle model's stages of create or receive; appraise and select; ingest; preservation action; store; access, use, and reuse; and transform.[6] That model is an ideal, and in practice some of the actions, such as store, are part of the entire process and not a distinct stage. As long as your program implementation is based on standards and best practice and works within your organizational context, your digital preservation program will be successful.

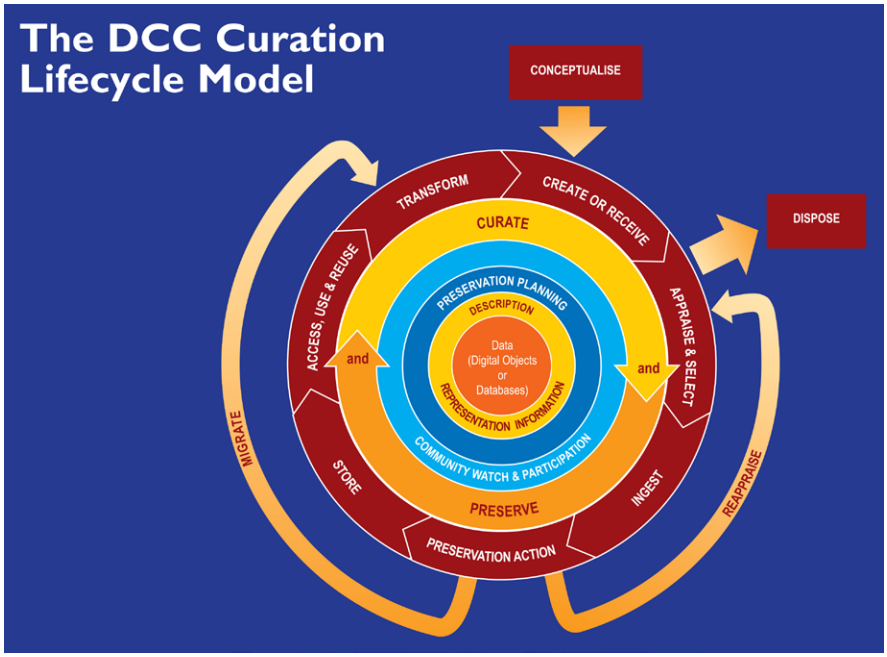

Figure 5.1
The digital curation life cycle model. Source: Digital Curation Centre, "DCC Curation Lifecycle Model," accessed June 5, 2019, www.dcc.ac.uk/resources/curation-lifecycle-model. CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/#). Image has been cropped.

The rest of this chapter will present a sample implementation strategy using open-source and proprietary tools. To be able to carry out the program as outlined requires a few pieces of hardware. The first is a computer with enough memory and processing power to use the digital preservation software tools. This computer, ideally, will rarely be connected to the internet in an effort to reduce the possibility of viruses being introduced to the computer and therefore to any digital materials being stabilized or processed on the computer. The second category of hardware are internal or external drives that allow you to access the various digital media items in your collection, such as a 3.5-inch floppy disc drive, a CD/DVD drive, and so forth. Lastly, you need some form of write blocker (hardware or software) that prevents you from inadvertently changing the metadata and digital content that you are trying to preserve. The sample strategy will not include storage recommendations because storage solutions need to be tailored to your specific needs, and there is excellent existing documentation and guidance already available for building preservation storage solutions.[7]

Accession

During the accessioning process, when you are discussing what will be donated to your organization and how, digital content requires you to deviate slightly from standard practice. First, it is advisable to modify your existing deed of gift to cover the preservation needs of the digital objects you will be receiving and the potential privacy breaches that can occur when the materials are stabilized. If one of your donors wishes to retain the rights to the intellectual property of the digital materials he or she is donating, you will need to negotiate the right to preserve these materials into the donor agreement. I also recommend that you include a section in your deed of gift that requires donors to give explicit permission for digital forensics activities, which enables donors to put limits on what your organization can and cannot do with the results of those activities. In an effort to support these choices, I suggest including a predetermined list of possible restrictions and redaction options. This additional section helps you to start the conversation about what you, as the processing archivist, or a very savvy potential user could find with access to a donor's digital materials so that donors have full knowledge of what they are agreeing to.

After you have negotiated with the donor over what will be donated and what you are allowed to do with the donation, ask the donor about the digital materials he or she is going to give you. I recommend you have a template checklist with all the potential questions for any kind of digital materials donation, and then you can tailor this template for each donor, subtracting those questions that do not apply. This checklist is most effective when you are able to sit down with donors and view the potential materials they would like to donate. That way you can go through the checklist and do a preliminary appraisal of the donation at the same time. You can also glean vital contextual clues about the organization of the digital materials and ask donors questions about why they chose to do things a certain way.[8] The answers to these questions will inform the arrangement decisions you make later on. You can also make suggestions to the donor about additional organization and potential migration from unique file formats to more common file types before the transfer to your institution occurs. Finally, you can decide how the digital content will be transferred from the donor to your organization, whether through digital media or through file transfer protocols.

If the digital materials are coming to you from inside your organization, you can bypass some steps. Instead of a donor agreement with a negotiation over intellectual property and informed consent, you have a transfer agreement that provides you with contextual information similar to the information you would receive from the donor checklist. The transfer agreement should also have instructions for internal personnel that describes how the digital content should be sent to you. It is important to include the transfer instructions because it is at the point of transfer where the digital content is extremely vulnerable to loss, of both intellectual content and technical metadata. Providing a strict procedure describing how transfers are supposed to occur can help reduce these risks.

Stabilize

Stabilization is the most important step of the entire process. Ideally, this will be done as soon as you receive digital material through an email attachment, on a flash drive, on a CD, and so forth. Stabilizing your digital content means that you are establishing a record of who you received the material from, in what condition it arrived, and the original metadata associated with the material (as best as you are able) and also establishing a baseline against which you will check, through time, to see if the digital material is ever modified in a way that affects its authenticity. After stabilization, you can safely let the digital materials sit in a monitored archival storage system until you have the time to finish the process. Throughout the entire procedure you should be documenting the actions you take for each file transfer or digital media transfer, either in your accession record or in a separate digital content log. This documentation should include the following information for each transfer or digital media item:

- Accession number
- Digital object identifier/transfer identifier

- Transfer type
- Date acquired
- Who donated the transfer
- Who received the transfer
- Media (if it came on digital media): format, manufacturer, model, age, condition, media label text
- Disk image (if a disk image is created): date created, disk image filename, disk image software used, notes, files exported (Y/N)
- Stabilization: virus scan (Y/N), checksums generated (Y/N), file format report generated (Y/N), personally identifiable information scanned for (Y/N), files moved to preservation storage (Y/N), notes
- Produce AIP: AIP created (Y/N), AIP saved in preservation storage (Y/N), date of transfer to storage, network storage location, notes

As early as you can in the stabilization workflow, ideally while the digital materials are still with the creator, but definitely before you transfer the files into your digital preservation processing system, run a virus check on the digital content being transferred. Your organization should be using some kind of virus protection software. Included in these software packages should be the ability to scan for viruses. I work almost exclusively in a Windows environment, so all I have to do is right-click on the directory I want to scan, chose the virus checker, and let it run. If your organization's software does not allow this, there are some open-source virus checkers available, or your information technology professionals may suggest one that they would prefer you use. In the event that a virus or malware is found in the material, contact your information technology department to see how they would like you to proceed. If you are working in a small shop, with no support from an IT department, put the files in quarantine and try again in ninety days after your virus protection software updates with all the new patches. In this case, quarantine may mean leaving the files on their digital media carriers or refusing to accept a file transfer. Depending on your virus protection software, it may take care of the entire quarantine and remediation process for you.

After virus checking, I recommend you set up a folder directory for the accession. The highest folder in the hierarchy should be named with the accession number. Within that, if you are going to create disk images, you will have three folders: disk_image, files, and metadata. If you don't create disk images, you will have only a files folder and a metadata folder. Within the disk_image and files folder, each individual transfer, digital media or otherwise, will get a folder named with the digital object identifier. In my case, the digital object identifier is the accession number and then a number starting at 001 and going up for however many digital media objects or file transfers are in the accession. As you move through the stabilization process, transfer files and save metadata into this folder structure. Example of a folder structure:

- 2007_038
  - 2007_038_disk_image
    - 2007_038_001_disk_image
    - 2007_038_002_disk_image
  - 2007_038_files
    - 2007_038_001_files
    - 2007_038_002_files
  - 2007_038_metdata

Now you are ready to transfer your digital materials to your digital preservation processing environment. If the transfer of digital content comes on digital media, you have two choices. You can create a disk image of the digital media, or you can do a direct copy of the files from the digital media to your digital preservation processing computer. Whether or not you create a disk image will be a matter of policy, donor agreement, and type of digital media. For instance, if I receive digital content on a flash drive or an optical disk, I very rarely create a disk image because the return on investment for these digital media does not often play out in my favor. If I receive a computer hard drive, I am much more likely to create a disk image because there is so much contextual information that can be retrieved and the potential for emulation. If the digital content does not come to you via digital media, your only option is via direct copying. If you decide to create a disk image, BitCurator is an open-source, community-supported software environment that has tools for creating, documenting, processing, and viewing disk images.[9]

The vast majority of the digital content transfers I facilitate are done by direct copy mechanisms. There are two tools that I recommend for this process, DataAccessioner and TeraCopy.[10] DataAccessioner allows you to supply metadata at the point of transfer for the accession, produces PREMIS metadata after running the File Information Tool Set (FITS) on the files, checks the fixity of files after they have been moved from the original source location into your processing environment, and does not alter the files' original metadata, such as last date modified or date of creation.[11] TeraCopy has free and for-purchase versions. The free version that I use will copy files without altering their internal metadata and performs a fixity check to make sure the files were not altered upon transfer.

After the files have been transferred or the disk image created and the files exported from the disk image for the entire accession, I suggest you create a file format report that is easy to read. This report serves two purposes: it is a record of the file formats in the accession so you can determine early on if special actions will be necessary immediately or in the near future to maintain access to these files and if you need

to research special software to gain access to the files. There a few different open-source tools available to do this job. You can compare them using POWRR Tool Grid or COPTR to find the one that works best for your organization.[12] I suggest that if the tool has a proprietary file format, you export the results as a comma-separated value (.csv) file type. A .csv opens well in many different spreadsheet and database software products.

The final step in the stabilization process is finding and documenting personally identifiable information (PII) and, for disk images only, extracting file system metadata. The BitCurator Environment includes the Bulk Extractor tool, which is what I use to do this. Bulk Extractor works for disk images and file sets. It generates reports on possible instances of PII and has a viewing tool that points you to where the PII is. If you have a disk image, I strongly encourage you to use the various reporting tools available in the BitCurator Environment to generate as much information about the original file system as possible.[13] At this point, you have a few choices. You can transfer the file directory for the accession into preservation storage as is until you have time to arrange and describe the files. You can make the file directory into a Submission Information Package and ingest it into your repository. Finally, you can go straight into appraising, arranging, and describing the content of the accession.

Appraise, Arrange, and Package

The appraise, arrange, and package step is very similar to the traditional version of archival processing when an archivist goes through the collection and determines what to keep, introduces a new arrangement for the files when necessary, and starts the process of creating an official description of the collection. One of the major differences between an analog collection and a digital collection is that digital collections can contain an exponentially higher number of individual "files" to go through, so you need different tools to get the job done.

I have found in practice, for hybrid collections, that you should process the analog materials before the digital. In this way, you get a feel for the creator's organizational style and start to have an idea of what you might find in the digital files. In some cases, your digital files will integrate seamlessly into your arrangement of the physical files, so each series, subseries, and so on will be a mix of physical and digital files. In other cases, you will have a series that is just computer files, but it is difficult to decide which way to go without having processed the physical materials. Processing the physical files will also help with the deduplication process.

My first step in appraisal of the digital content in an accession is to look at a visualization of the content. The two tools I use to do this are WinDirStat and TreeSize.[14] I use WinDirStat to get a quick overview of the different types of files in the accession or as a teaching tool to help internal content creators understand what they are producing and how to find out what is taking up all of their computer's memory. More often, I use the professional version of TreeSize. Not only does it show me a visual representation of the different types of file formats in a collection, but it also contains tools for deduplication of files and for finding and removing temporary files, internet files, and more. This can dramatically reduce the number of files I will eventually need to arrange.

At this point you can choose to leave the files as they are, or you can go further in the arrangement of the files. What you do will be determined by your organization's processing workflow because the decision-making factors for digital materials at this point are rarely different from those for physical materials. Generally, when I make the decision to not keep the original arrangement of digital files, it means that I am matching the arrangement of the physical files or that the collection has no discernable useful organization and therefore it would be too difficult for users to navigate in its original state. If you decide to arrange the files into a new folder structure, I recommend that you create the new folder structure first and then move the files into it. At this point you may also be renaming files individually, but this is not necessary.

After you have settled on the final arrangement of the files, you can either use a series of tools to, at the bare minimum, strip file names of special characters and normalize files for preservation and access, or you can use a tool like Archivematica, which will automate the SIP to AIP and DIP process for you.[15] Archivematica uses a series of customizable microservices that document your digital collections, perform preservation actions such as assigning a unique identifier to each digital file, remove special characters from file names, and so much more. Archivematica will also and transform the files into preservation and access versions of the original digital files when it is able to do so. You can either pay for a hosted version of the software or try and maintain your own instance of the software. Be aware that in practice Archivematica takes a lot of technical knowledge and can require a lot of maintenance. Archivematica can integrate into ArchivesSpace, Archivists' Toolkit, AToM, DSpace, or your own preservation storage. If you do not use Archivematica, there are a series of automated file renaming, file migration and normalization, and packaging tools on COPTR for you to use. From here, you can move your AIPs into preservation storage and your DIPs into your access storage environment.

A final piece of the preservation puzzle is fixity monitoring. It is not enough to create checksums of all of your digital materials upon transfer. You also need to check to make sure that each checksum does not

change over time due to an accident, malicious activity, or simple bit rot. It is impossible to do this by eye, especially if your program includes millions of files. Instead, I recommend that you use fixity monitoring software that will run on a schedule and notify you of any changes. It is possible that your information technology department already uses such a service. If so, communicate with them to get access to the fixity report. If not, COPTR has a couple of options of open-source fixity monitoring tools for you to choose from.

Access

Access to digital collections for your users may take many different avenues depending on your institution's resources. It is just as valid to have users request access to the digital materials in a collection as described in your finding aid and provide them with a link to a shared folder through an online drop box as it is for users to have immediate access to digital content through a digital library or repository system. In some cases, users may access your content only from a reading room computer. The most likely solution is a combination of all of these, depending on your resources and the specific restrictions, if they exist, for each collection. What is essential is that users *have* access and there be clear documentation describing how users may or may not gain access to collections and why. At the bare minimum, a fully processed digital collection will have a finding aid or catalog record with information about how to request access to the materials.

Maintenance

An essential piece of digital preservation program implementation is maintenance. As a digital preservation practitioner, you should try to pay attention to the wider world of digital preservation literature and tool development. As new tools and services become available, evaluate them against what you are already doing. If they are an improvement, determine if your current resources would allow you to integrate a new tool or service into your existing system. If not, it may be time to create an updated business case to ask your organization's leaders for additional resources. Another part of maintenance is planning for the inevitable replacement of your existing software and hardware solutions. Hopefully, some of the burden of replacement will be shared by your information technology division. If you do not have an information technology division, it may be better to plan for a transition into storage as a service, such as cloud storage, so that you do not have to maintain your own storage infrastructure. Finally, do your own personal maintenance—attend digital preservation conferences, workshops, and webinars when you are able to. By increasing your own knowledge of digital preservation, you will create more efficient workflows and be able to modify existing tools so they work better for your organization. You will also become a stronger advocate for your digital preservation program and be better able to introduce and maintain collaborations with other digital preservation programs.

Notes

1. Digital Curation Centre, "DCC Curation Lifecycle Model," accessed June 5, 2019, www.dcc.ac.uk/resources/curation-lifecycle-model.
2. Example end-to-end preservation systems include Preservica (https://preservica.com/) and Ex Libris Rosetta (https://www.proquest.com/products-services/Ex-Libris-Rosetta.html).
3. "Tool Grid," Digital POWRR: Digital Preservation Research, 2013, https://digitalpowrr.niu.edu/digital-preservation-101/tool-grid/; COPTR homepage, accessed June 5, 2019, http://coptr.digipres.org/Main_Page.
4. Brian Lavoie, *The Open Archival Information System (OAIS) Reference Model: Introductory Guide*, 2nd ed., DPC Technology Watch Series (Glasgow, Scotland: Digital Preservation Coalition, October 1, 2014), https://doi.org/10.7207/twr14-02.
5. Mukurtu CMS homepage, accessed June 6, 2019, http://mukurtu.org.
6. Digital Curation Centre, "DCC Curation Lifecycle Model."
7. Christopher J. Prom, Erin O'Meara, and Kate Stratton, *Digital Preservation Essentials* (Chicago: Society of American Archivists, 2016); Digital Preservation Coalition, "Digital Preservation Handbook," 2015, https://www.dpconline.org/handbook.
8. Melissa Watterworth Batt, "Donor Survey DRAFT," Electronic Records Committee, Congressional Papers Section, Society of American Archivists, October 31, 2012, https://cprerc.files.wordpress.com/2015/08/sample-donor-survey_dodd-center_draft.pdf.
9. BitCurator homepage, accessed June 6, 2019, http://bitcurator.net.
10. DataAccessioner homepage, accessed June 6, 2019, http://dataaccessioner.org; "TeraCopy for Windows," Code Sector, accessed June 6, 2019, https://www.codesector.com/teracopy.
11. FITS is a tool developed and maintained by Harvard that "identifies, validates and extracts technical metadata for a wide range of file formats." File Information Tool Set (FITS) homepage, accessed June 6, 2019, https://projects.iq.harvard.edu/fits/home.
12. "Tool Grid"; COPTR homepage.
13. BitCurator homepage.
14. WinDirStat homepage, last updated November 12, 2018, https://windirstat.net; TreeSize Professional webpage, JAM Software, accessed June, 21, 2019, https://www.jam-software.com/treesize.
15. Archivematica homepage, accessed June 6, 2019, https://www.archivematica.org/en.